NON-FLAT CLUSTERING WHITH ALPHA-DIVERGENCES

Olivier Schwander[⋆†]*, Frank Nielsen*[⋆‡]

[⋆] École Polytechnique, Palaiseau, France
[†] ÉNS Cachan, France
[‡] Sony CSL, Tokyo, Japan

ABSTRACT

The scope of the well-known k-means algorithm has been broadly extended with some recent results: first, the k-means++ initialization method gives some approximation guarantees; second, the Bregman k-means algorithm generalizes the classical algorithm to the large family of Bregman divergences. The Bregman seeding framework combines approximation guarantees with Bregman divergences. We present here an extension of the k-means algorithm using the family of α-divergences. With the framework for representational Bregman divergences, we show that an α-divergence based k-means algorithm can be designed. We present preliminary experiments for clustering and image segmentation applications. Since α-divergences are the natural divergences for constant curvature spaces, these experiments are expected to give information on the structure of the data.

Index Terms— $alpha$-divergence, clustering, information geometry, k-means

1. INTRODUCTION AND PRIOR WORK

Originally restricted to the Euclidean distance, the well-known Lloyd's k-means algorithm [1] has been extended to a larger family of distortion measures, the Bregman divergences (which contains the Euclidean distance) by Banerjee *et al.* [2]. Even if the output of the classical k-means is a local minimizer of the loss function, it is known that the result can be arbitrary far from the optimum and that the quality of the clusters is heavily dependent on the initial seeds chosen to represent the clusters: Arthur and Vassilvitskii [3] presented the k-means++ initialization method (also known as squared Euclidean seeding) which guaranties to have a $O(\log k)$-approximation of the optimal output by carefully choosing the initial seed of the k clusters. The two results have been unified recently by Nock *et al.* [4] to provide a Bregman clustering algorithm with approximation guarantees.

We introduce here an extension of the previous results on the family of α-divergences. Such an extension is an important complement of the Bregman clustering since the Bregman divergences are the canonical divergences for dually flat spaces and the α-divergences are the canonical divergences for constant curvature spaces. This is not the first attempt to take into account the specificities of constant curvature spaces, Dhillon and Modha [5] introduced the spherical k-means algorithm which use a cosine similarity but our algorithm is more general since it allow a more precise choice of parameters and is not limited to positive curvature spaces.

This new work is made possible by the representational Bregman divergences framework introduced by Nielsen and Nock [6]. In this framework, each α-divergence is seen as the Bregman divergence generated by some strictly convex and differentiable function (the Bregman generator) acting on an adapted representation function.

Section 2 recalls some definitions and presents the representation Bregman divergences framework. Section 3 presents the clustering algorithm itself. Section 4 presents few preliminary experiments.

2. REPRESENTATIONAL BREGMAN DIVERGENCES

2.1. Definitions

Invariance and information monotonicity of α-divergences

We recall the definition of α-divergences [7] that are defined on positive arrays (unnormalized discrete probabilities) for $\alpha \in \mathbf{R}$ as:

$$D_\alpha(p\|q) = \begin{cases} \sum_{i=1}^d \frac{4}{1-\alpha^2}\left(\frac{1-\alpha}{2}p_i + \frac{1+\alpha}{2}q_i - p_i^{\frac{1-\alpha}{2}} q_i^{\frac{1+\alpha}{2}}\right) & \text{if } \alpha \neq \pm 1 \\ \sum_{i=1}^d p_i \log \frac{p_i}{q_i} + q_i - p_i = \mathrm{KL}(p\|q) & \text{if } \alpha = -1 \\ \sum_{i=1}^d p_i \log \frac{q_i}{p_i} + p_i - q_i = \mathrm{KL}(q\|p) & \text{if } \alpha = 1 \end{cases} \quad (1)$$

This is all the more important that in the heart of many computer vision application, we deal with histograms (e.g., SIFT descriptors [8], GIST descriptors [9]) that are considered as multinomial probability distributions. Therefore, we need a distribution measure D to calculate the dissimilarity of multinomials $D(p(x;\theta_p)||p(x;\theta_q))$ where θ_p and θ_q

are the histogram distributions. Symmetrized α-divergences $S_\alpha(p,q) = \frac{1}{2}(D_\alpha(p\|q) + D_\alpha(q\|p))$ belong to Csiszár's f-divergences and therefore retain the information monotonicity property.

From the pioneering work of Chentsov [10], it is known that the Fisher-Rao riemannian geometry (with the induced Levi-Civita connection) and the α-connections are the *only* differential geometric structures that preserve the measure of probability distributions by reparameterization. We consider the α-divergences that are a proper sub-class of Csiszár f-divergences that satisfy both reparameterization invariance (i.e., $D(p(x;\theta_p)||p(x;\theta_q)) = D(p(x;\lambda_p)||p(x;\lambda_q))$ for $\lambda_x = f(\theta_x)$ where f is a bijective mapping) and information monotonicity [11]: $D(p(x;\theta_p)||p(x;\theta_q)) \geq D(p(x;\theta'_p)||p(x;\theta'_q))$ for θ' a coarser partition of the histogram. That is, if we merge bins θ into coarser histograms θ', the distance measure should be less than the distance by considering the higher-resolution histograms.

Bregman divergences

Given a strictly convex and differentiable function $F : \mathbf{R}^d \to \mathbf{R}$, we define the Bregman divergence associated with the generator F as:

$$B_F(p\|q) = F(p) - F(q) - \langle p - q, \nabla F(q)\rangle \quad (2)$$

The generator $F(x) = x^\top x = \sum_{i=1}^d x_i^2$ yields to the squared Euclidean distance. Using the Shannon negative entropy ($F(x) = \sum_{i=1}^d x_i \log x_i$) we get the well-known Kullbach-Leibler (KL) divergence.

2.2. Representation function

Nielsen and Nock [6] showed that α-divergences (but also β-divergences [12]) are representational Bregman divergences in disguise. Let's consider *decomposable* Bregman divergences:

$$B_F(p\|q) = \sum_{i=0}^d B_F(p_i\|q_i) \quad (3)$$

With a slight abuse of notation, we denote its separable generator F as $F(x) = \sum_{i=0}^d F(x_i)$. We call representation function a strictly monotonous function k that introduces a (possibly non-linear) coordinate system $x_i = k(s_i)$ where each s_i comes from the source coordinate system. This mapping is bijective since k is strictly monotonous and $s_i = k^{-1}(x_i)$. We have the following Bregman generator:

$$U(x) = \sum_{i=1}^d U(x_i) = \sum_{i=1}^d U(k(s_i)) = F(s) \quad (4)$$

where $F = U \circ k$.

The class of α-divergences are representational Bregman divergences for

$$U_\alpha(x) = \frac{2}{1+\alpha}\left(\frac{1-\alpha}{2}x\right)^{\frac{2}{1-\alpha}}, \quad k_\alpha(x) = \frac{2}{2-\alpha}x^{\frac{1-\alpha}{2}} \quad (5)$$

Notice it turns out that F may not be strictly convex [6] ($U_\alpha \circ k_\alpha$ is linear) although U is always *strictly* convex.

2.3. Centroids

Like (most of) the Bregman divergences, α-divergences are not symmetrical. This yields to two different ways of defining centroids: the left-sided centroid c^L and the right-sided centroid c^R:

$$\begin{array}{l} c^R = \arg\min_{c\in\mathcal{X}} \sum_{i=1}^n B_{U,k}(p_i\|c) \\ c^L = \arg\min_{c\in\mathcal{X}} \sum_{i=1}^n B_{U,k}(c\|p_i) \end{array} \quad (6)$$

Closed-form formulas are given in [6]:

$$\begin{array}{l} c^R = k^{-1}\left(\sum_{i=1}^n k(p_i)\right) \\ c^L = k^{-1}\left(\nabla U^*\left(\sum_i \nabla U(k(p_i))\right)\right) \end{array} \quad (7)$$

where U^* is the Legendre convex conjugate of U, see [13]. In the particular case of α-divergences, we obtain:

$$\begin{array}{l} c^R = n^{-\frac{2}{1-\alpha}}\left(\sum p_i^{\frac{1-\alpha}{2}}\right)^{\frac{2}{1-\alpha}} \\ c^L = n^{-\frac{2}{1+\alpha}}\left(\sum p_i^{\frac{1+\alpha}{2}}\right)^{\frac{2}{1+\alpha}} \end{array} \quad (8)$$

3. CLUSTERING WITH REPRESENTATION FUNCTIONS

3.1. k-means algorithm

Banerjee *et al.* [2] showed that the classical clustering algorithm k-means generalizes to and only to Bregman divergences. Using the representational framework of section 2.2, we extend their algorithm to the α-divergences by plugging the representation function: this is algorithm 1. (Symmetrized α-divergences S_α are handled implicitly by two potential functions, similarly to [4].)

We can establish the following proposition:

Proposition 1. *The representational Bregman k-means (algorithm 1) monotonically decreases the loss function $L_{U,k}$.*

Algorithm 1 Representational Bregman k-means

Require: A set $\mathcal{X}$ of n points x_i of $\mathbf{R}^d$, a number of clusters k, a representational divergence $B_{U,k}$
Choose k points μ_i (with some seeding method)
repeat
{Assignment step}
Set $\mathcal{X}_h \leftarrow \emptyset$ for $1 \leq h \leq k$
for $i = 1$ to n **do**
$h \leftarrow \arg\min_{h'} B_{U,k}(x_i || \mu_{h'})$
Add x_i to $\mathcal{X}_h$
end for
{Relocation step}
for $h = 1$ to k **do**
$\mu_h \leftarrow k^{-1} \left(\sum_{i=1}^{n} k(x_i)\right)$
end for
until convergence
Return $\{\mu_1, \ldots, \mu_k\}$

Proof.

$$L_{U,k}^{(t)} = \sum_{h=1}^{k} \sum_{x_i \in \mathcal{X}_h^{(t)}} B_{U,k}(x_i, \mu_h^{(t)}) \quad (9)$$

$$\geq \sum_{h=1}^{k} \sum_{x_i \in \mathcal{X}_h^{(t)}} B_{U,k}(x_i, \mu_{h^\star(x_i)}) \quad (10)$$

$$\geq \sum_{h=1}^{k} \sum_{x_i \in \mathcal{X}_h^{(t+1)}} B_{U,k}(x_i, \mu_{h^{(t+1)}}) \quad (11)$$

where $h^\star(x_i) = \arg\min_{h'} B_{U,k}(x_i || \mu_{h'})$

The inequality (10) comes from the assignment step: each point is re-assigned to the cluster with the nearest centroid.

The inequality (11) comes from the re-estimation step: the centroid of each cluster is recomputed, reducing the cost of each cluster. □

Thus, the Representational Bregman k-means algorithm gives a partition which is locally optimal.

3.2. Representational seeding

The Bregman seeding method presented in [4] can be generalized in a straightforward way in order to get the same approximation guarantees as in in k-means++ [3]. Algorithm 2 describes how to seed the initial cluster. The proof is omitted but can be drawn directly from the [4] one.

4. EXPERIMENTS

4.1. Clustering

We tried our algorithm on one of the most natural space with a positive constant curvature: the Earth. We clusterized a set of 61 cities divided in continents and described by their coordinates in a 3-dimensional space. The resulting clusters were evaluated in terms of Normalized Mutual information [14]. We saw in our experiments that the α parameter has a little impact on clustering performance (figure 1). We explain this by the fact that the dataset used is easy and does not need a careful use of its geometry.

Algorithm 2 Representational seeding

Require: A set $\mathcal{X}$ of n points x_i of $\mathbf{R}^d$, a number of clusters k, a representational divergence $B_{U,k}$
Set $\mathcal{C} \leftarrow x$ with x chosen uniformly at random in $\mathcal{X}$
for $i = 1$ to $k - 1$ **do**
Choose x in X with probability

$$p(x) = \frac{B_{U,k}(x, c_x)}{\sum_{y \in A} B_{U,k}(y, c_x)}$$

where $c_x = \arg\min_{z \in \mathcal{C}} B_{U,k}(x, z)$
Add x to $\mathcal{C}$
end for
return $\mathcal{C}$

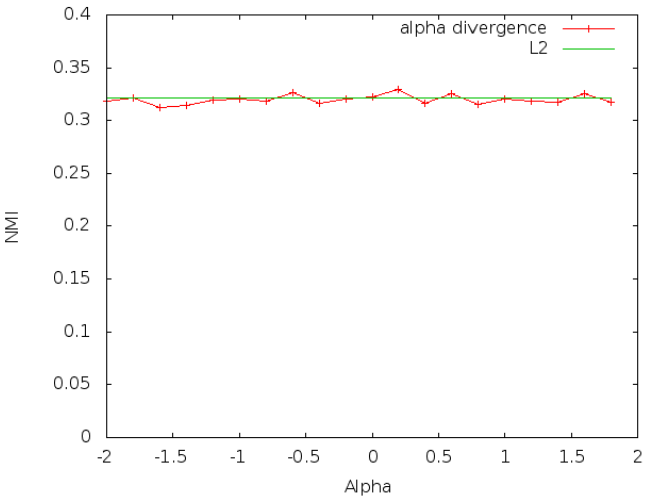

Fig. 1. Clustering results on cities dataset

4.2. Segmentation

We present results on segmentation application: each image is seen as a set of points in the five dimensional *RGBXY* space. The quality of the segmentation is evaluated visually.

The parameter α was moved from -1000 to 1000 with a thinner step for values near 0. We use 4, 8 and 16 different clusters for segmentation.

As one can see of the examples images of the figure 2 the segmentation little depends on the α-parameter.

Fig. 2. Segmentation for α in {-1000, -100, -10, -1, 0, 1, 10, 100, 1000}, for 8 clusters. All these images differ only by few pixels and are very similar to the one obtained with the Kullbach-Leibler divergence ($\alpha = \pm 1$).

5. CONCLUSION

The Bregman k-means algorithm presented in this paper extends well to the case of the α-divergences. This extension opens a new field of distortion measures for clustering related applications. Even if the intuition should be that a careful choice of the α could lead to improvements, this was not the case on the examples of clustering and segmentation we showed: since it was made only on one simple dataset and on two images, it can not be used to draw a general conclusion but it will need further work to better understand this result.

We intend to extend our study to synthetic datasets in order to validate the link between α-divergences and constant curvature spaces and to real datasets in order to understand at which extent it can benefit from α-clustering. It would be particularly interesting to discover the curvature of some famous image descriptors (such as GIST or SIFT).

6. REFERENCES

[1] S.P. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[2] A. Banerjee, S. Merugu, I.S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *The Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, 2005.

[3] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2007, p. 1035.

[4] R. Nock, P. Luosto, and J. Kivinen, "Mixed Bregman clustering with approximation guarantees," Springer.

[5] I.S. Dhillon and D.S. Modha, "Concept decompositions for large sparse text data using clustering," *Machine Learning*, vol. 42, no. 1, pp. 143–175, 2001.

[6] F. Nielsen and R Nock, "The dual Voronoi diagrams with respect to representational Bregman divergences," in *International Symposium on Voronoi Diagrams (ISVD)*, 2009.

[7] S.I. Amari and H. Nagaoka, *Methods of information geometry*, AMS Bookstore, 2007.

[8] D.G. Lowe, "Object recognition from local scale-invariant features," in *International Conference on Computer Vision*, 1999, vol. 2, pp. 1150–1157.

[9] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," *Progress in Brain Research*, vol. 155, pp. 23, 2006.

[10] N.N. Chentsov, *Statistical Decision Rules and Optimal Inferences*, vol. 53, Trans. of Math. Monog., 1982.

[11] Imre Csiszár, "Axiomatic characterizations of information measures," *Entropy*, vol. 10, no. 3, pp. 261–273, 2008.

[12] M. Mihoko and S. Eguchi, "Robust blind source separation by beta divergence," *Neural computation*, vol. 14, no. 8, pp. 1859–1886, 2002.

[13] F. Nielsen, J.D. Boissonnat, and R. Nock, "Bregman Voronoi diagrams: Properties, algorithms and applications," *Institut National de Recherche en Informatique et en Automatique (INRIA Sophia Antipolis), Research Report*, vol. 6154, 2007.

[14] A. Strehl and J. Ghosh, "Cluster Ensembles–A Knowledge Reuse Framework for Combining Multiple Partitions," *Journal of Machine Learning Research*, vol. 3, pp. 583–617, 2002.